The UX Construct – Does the Usage Context Influence the Outcome of User Experience Evaluations?

Andreas Sonderegger, Andreas Uebelbacher, Jürgen Sauer

The UX construct – does the usage context influence the outcome of user experience evaluations?

Andreas Sonderegger[1], Andreas Uebelbacher[2] and Jürgen Sauer[3]

[1] EPFL+ECAL Lab, EPFL Lausanne, Switzerland
[2] Access for All, Access for All, Zürich, Switzerland
[3] Department of Psychology, University of Fribourg, Fribourg, Switzerland

Abstract. How are different measures of user experience (UX) related to each other? And does it differ if a technological device is used for work or leisure with regard to UX? In the present study, the influence of context factors (i.e. usage domain) on the outcomes of UX tests is examined. Using a 2 x 2 experimental design, in addition to usage domain (work vs. leisure), system usability was manipulated (normal vs. delayed response time). Sixty participants completed various tasks with a mobile internet application. Performance indicators and subjective indicators of UX were recorded (e.g. emotion, perceived usability, and task load). Interestingly, results indicated little evidence for an influence of usage context on UX. System usability showed the expected effects on performance and on user emotion, whereas no influence on perceived usability was observed. In addition, the correlations between the different measures of UX were rather low, indicating that it is advisable to assess UX by distinct dimensions. Implications of these results for practice and research are discussed.

Keywords: user experience, UX, UX test, UX measures, system response time, usage domain, perceived usability, user performance.

1 Introduction

There is a long tradition in psychology to conceive behaviour as being highly dependent on context [1]. This also applies to the field of human-computer interaction (HCI), where context is considered one of the main determinants of user behaviour when operating interactive systems [2]. The importance of the concept is reiterated by the fact that it is also part of ISO FDIS 9241-210, which defines context as covering users, tasks and equipment, and the specific social and physical environment in which a product is used [3]. In this regard, a product's usability can only be evaluated taking into account the context in which the product will be used [4]. In other words, a system that may be usable in one context may not be in another. Context factors however play not only an important role in the evaluation of usability of a system but may influence the entire experience of a user. User experience (UX) can be considered an extension of the usability construct, which has a focus on user cognition and performance, by adopting a more holistic approach focusing on user emotions and considering the experience of a

user interacting with an interface in its entirety. The assumption of such a holistic approach to the construct suggests to individually assess and report the various facets of experience in UX evaluations.

While users, tasks and equipment are routinely specified in UX studies, the environmental aspects of context are rarely considered in practice and research [5, 6]. However, empirical work has provided evidence that a number of context factors might influence the outcomes of UX tests, such as lab vs. field set-up [7], observer presence [8, 9], or the use of electronic recording equipment [10]. One important characteristic of the usage environment which received little attention in previous research is the influence of the usage domains of leisure and work context on outcomes of UX evaluations.

2 Current state of research

2.1 The UX construct

Up into the 1980s, most people experienced interactive technology almost exclusively in the workplace [11]. Since then, information technology became an integral part of our daily lives and increasingly pervades all societal activities (e.g. personal computers have reached people's homes, and mobile phones have become mobile computing devices). Today, the pervasiveness of interactive technology in all areas of people's lives, including leisure, is a reality [12]. In this process, the distinction between technology for work and technology for leisure use has become increasingly fuzzy, as devices often cannot be described anymore as clearly being one or the other. Many of today's technical devices are dual-domain products [13], as they can equally be used in a work context as in a leisure context (e.g. mobile phones and laptops).

Since research in HCI traditionally concentrated on interactive technology in the work domain, the discipline was primarily performance oriented, with the goal to provide highly usable interfaces to increase efficiency at work [14]. This is reflected in the usability definition of the International Standardisation Organisation (ISO), which describes the concept in terms of efficiency and effectiveness (and satisfaction) with which a user can accomplish certain tasks with a system [3]. As a result of a recent shift in the domain towards a less functional and more experiential approach, the notion of UX has gained an increased interest in research and practice. Despite its popularity, the term UX is often criticised for being ill defined and elusive [1, 15]. While some theorists adopted a rather holistic approach describing UX as the totality of actions, sensations, thinking, feelings and meaning-making of a person in a specific situational context [e.g. 16, 17], others attempted to be more concise and focus on emotions or affective states when UX is described [e.g. 18–20]. This piece of research follows a holistic view defining UX as an umbrella construct which encompasses the entire experiential space of person interacting with a system. In this respect, UX enlarges the mere functional concept of usability (e.g. effectiveness, efficiency, subjective appraisal) by explicitly encompassing a broad range of other experiential components (c.f. Fig. 1). This holistic view of UX implies however the difficulty of a meaningful and effective assessment of the construct. This is because it might be rather complicated to measure

everything a user experiences when interacting with a system. Furthermore, the holistic view of UX makes it difficult to estimate one exact UX score since it is not clear how measures of affective sates are to be combined with indicators of performance and evaluations of satisfaction and workload. In this regard, we suggest to define the most relevant experiential facets or dimensions for each UX evaluation individually based on the specific needs and requirements of the project. These facets or dimensions should then be assessed and reported individually (as suggested to some extent by [21] in their UX scale, see also [6, 15, 22]).

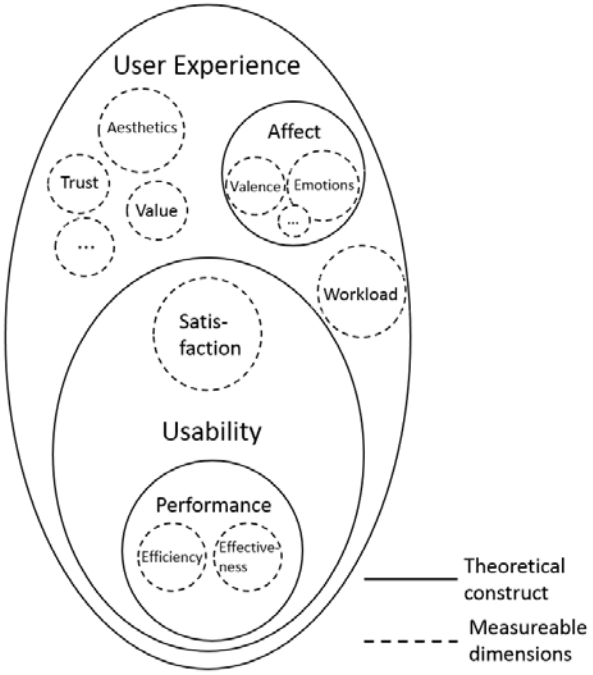

Fig. 1. The user experience construct and its components

2.2 Work vs. leisure domain

As a result of the shift from usability to UX, research in HCI has started to increasingly evaluate leisure-oriented technology, such as portable digital audio players [23] or video games [24]. For dual-domain products, however, both domains are equally relevant and different requirements may result from these contexts, which might need to be considered during product evaluation. In order to identify the domain-specific requirements, the differences between work and leisure need to be analysed. There is one previous study that empirically compared work and leisure domains but found little differences between them [13]. However, this may be due to the fact that in addition to usage domain, product aesthetics was manipulated as a second factor. It was assumed that aesthetics would play a more important role in a leisure context than at work but

this was not confirmed. Since system usability is a more direct determinant of effectiveness and efficiency of use than aesthetics, its influence in different usage domains is worth examining.

In addition to the lack of empirical research in that field, there have also been difficulties in establishing a clear theoretical distinction between work and leisure so that no widely accepted definition of the concepts has yet been proposed. Three approaches to distinguish between the two domains have been discussed [25]. (a) The purely time-based or 'residual' definition of leisure is most commonly used. According to this approach, leisure is when people do not do paid or unpaid work, do not complete personal chores, and do not fulfil obligations. (b) An activity-based approach distinguishes between work and leisure by means of specific behaviours people show in each domain. (c) The third approach conceives work and leisure by the attitudes people have towards their activities. Beatty and Torbert [25] argue that the third approach is considered to be most promising to distinguish leisure from other domains, and there is also some empirical evidence in support of this approach. Several studies confirmed that people described work in terms of goal-directed and performance-oriented behaviour and connected with external rewards while leisure was associated with intrinsic satisfaction, enjoyment, novelty and relaxation [26, 27].

The distinction between work and leisure domain according to this third approach allows for a more precise definition of domain-specific requirements for interactive technology. Since users perceive a work context as more goal- and performance-oriented than a leisure context, usability (e.g. effectiveness and efficiency of task completion) might be perceived by users as a more important requirement in a work than in a leisure context.

2.3 Response time as a facet of system usability

One aspect of system usability which previous research has shown to be directly relevant for various outcome variables is system response time (SRT). SRT is defined as the time it takes from a user input to the moment the system starts to display the response [28]. Although delayed SRT are less of a problem with today's much increased processing power, delays may still be a problem in human-computer interaction [29, 30].

Negative effects of SRT delays have been shown at several levels. First, there is evidence that response time delays have a negative effect on user satisfaction with a system [31–34]. Systems with delayed responses are generally perceived as being less usable and more strenuous to operate, which also extends to web pages with long download times being judged to be less interesting [35]. Second, user performance has been shown to be impaired by SRT delays [28, 29, 36, 37]. Third, system delays have resulted in impaired psychophysiological well-being, increased anxiety, frustration and stress, and were even found to reduce job satisfaction [31, 38–41].

Various moderators of the effects of system response delays have been identified in the context of internet usage, such as webpage properties [42] user expectations [43], and processing information displays such as progress bars [44]. For example, users

were less willing to accept download delays when websites were highly graphical compared to plain text documents [42]. It also emerged that information about the duration of the download had a positive effect on user evaluation [43], and progress bars as delay indicators performed best in terms of user preference and acceptability of the waiting time [44]. To our knowledge, only one study has researched SRT in a work context [31]. Conducting a field study in a large telephone circuit utility observing professionals in their work domain, the authors reported that increased SRT not only impaired performance but also system evaluation and even job satisfaction. To our knowledge, no study investigated SRT delays in a leisure context so far.

2.4 The present study

The main goal of the present study was to investigate the requirements that result from the two domains of work and leisure with respect to UX design and evaluation. For this purpose, a UX test was conducted in which the two domains of work and leisure were experimentally modelled. The two types of testing context were created by a combination of various experimental manipulations such as lab design (office vs. living room), task wording (work related vs. leisure related) and a priming task which directed participants' attention towards their own work or leisure activities, respectively. As a second independent variable, system usability was manipulated through SRT delays.

As a test system, an internet site was specifically set up for the experiment, which was designed to offer realistic tasks for both contexts. Care was taken that the tasks for the two experimental conditions were comparable in terms of mental demands but only differed in type of context. The tasks used were information search tasks that required navigating through various levels of a menu hierarchy.

Typical measures for UX evaluation were recorded. Task performance was assessed by task completion rate, page inspection time, and efficiency of task completion. Self-report data was collected for emotion, task load and perceived usability.

Our hypotheses were as follows: (a) Test participants in the work context perform better and report higher perceived task demands than those in the leisure context, since the work context is perceived as more goal- and performance-oriented. (b) Performance is lower when working in the condition with low usability compared to working with high usability. (c) Perceived usability of the system and emotional reactions are less positive in the low usability condition, since the reduced system usability is reflected in participants' evaluation and emotion. (d) At work, low usability causes a stronger decrease in perceived usability and in emotion than in the leisure context, since the negative impact of system delay on performance is perceived as more relevant in the goal- and performance-oriented work context.

3 Method

3.1 Participants

The sample of the experiment consisted of 60 participants, aged between 19 and 44 years (M = 22.6 yrs; SD = 3.3), the majority of which were female (60.3%). Participants

were recruited among students at the University of Fribourg (Switzerland), and it was made sure that none of them had previous usage experience with the specific mobile phone model employed in the experiment. To motivate participants to take part in the study, they could enter a prize draw (worth 50 $).

3.2 Experimental design

A 2 x 2 between-subjects design was used to investigate the two independent variables usage domain and usability. Usage domain was varied at two levels (work vs. leisure context), and so was usability (high vs. delayed system response time).

3.3 Measures and instruments

Performance
The following three measures of user performance were recorded: (a) task completion rate (percentage of successfully completed tasks); (b) page inspection time (time a user stays on a page); (c) efficiency of task completion (minimum number of interactions needed for task completion divided by actual number of interactions). Participants were allowed to work on each task for a maximum of 5 min, after which a task was recorded as failed and participants moved on to the next task. All analyses of performance data took into account the shorter overall time participants had available in the delay condition (i.e. delay time was deducted from task completion time).

Affective state
The PANAS scale ('Positive and Negative Affect Schedule' [45]) was used to measure short-term changes in affective states before and after task completion. The scale allows the assessment of two independent dimensions of affect: positive and negative affect. It was shown to have good psychometric properties (Cronbach's $\alpha = 0.84$). The scale uses 20 adjectives to describe different affective states (e.g. 'interested', 'exciting', 'strong'), for which the intensity is rated on a 5-point Likert scale ('very slightly or not at all', 'a little', 'moderately', 'quite a bit', 'extremely').

Task load
To assess task load the well-established NASA task load index (TLX) was used [46]. It measures the six dimensions of task load: mental demands, physical demands, temporal demands, performance, effort and frustration. In the subsequent analysis, each dimension was given the same weight. Based on our data, psychometric properties were shown to be satisfactory for the translated scale (Cronbach's $\alpha = 0.72$).

Perceived usability
Perceived usability of the test system was measured by two instruments. First, we used a 100mm visual analogue scale to measure an overall evaluation of perceived usability ('This website is usable') [8]. The use of one-item scales to evaluate technical systems

was found to be appropriate, as other work has shown [e.g. 47]. Second, the PSSUQ ('Post Study System Usability Questionnaire') [48] was applied, which was slightly modified to be relevant for the test system in question (the term 'system' was replaced by 'software' to make sure only the software and not the device was judged). The scale consists of 19 items and uses a 7-point Likert scale (ranging from 'strongly agree' to 'strongly disagree'). The questionnaire was developed for usage in usability tests in a lab setting, and the author [48] reports very good psychometric properties (Cronbach's $\alpha > 0.90$).

Previous mobile phone experience

Previous mobile phone experience was assessed by a visual analogue scale on which participants reported an intermediate self-rated mobile phone experience of 5.0 on a scale ranging from 0 to 10 (labelled 'not experienced' and 'very experienced'). They indicated using their devices 12.6 times on average during a day. Mobile phone experience and daily usage were used as covariates in the analysis.

3.4 Materials: Mobile phone, server and software

The test device was a Motorola Android smartphone. The web application was running on a server software XAMPP. In the delay condition, a PHP script was running on the server and generated a random system response delay of between 0s and 3s (1.3s on average) whenever a new page was requested. These system delays were chosen based on pre-tests which showed that changing intervals were perceived as more disturbing than constant (and hence predictable) ones. In addition, the pilot tests have shown that latencies of more than 4 seconds were not considered realistic. A server log recorded the pages viewed, the time at which the page was accessed, the duration during which the page was displayed and the size of the delay.

The web application used for task completion was specifically set up for the experiment. It consisted of a tourist guide for a large European city containing a hierarchical navigation system, offering a number of categories at each level and detailed pages at the deepest level. Navigation options were 'return to the previous page', 'return directly to the home page', or selecting one of the displayed categories. Scrolling was necessary for some of the pages, which had a larger number of categories than the screen could display. Category labels were deliberately named such that it was not always obvious in which the target page would be found so that a trial and error approach to target search became necessary (e.g. a specific Asian restaurant was located under the category 'Japanese', while other categories available included 'Asian', 'Chinese', 'German', 'Greek', 'Indonesian', and 'Italian'). A message on the target page stated clearly that the task had been solved and requested that the user directly went back to the home page.

3.5 Procedure

Participants were randomly assigned to one of the four testing conditions. The testing sessions were conducted in a usability laboratory at the University of Fribourg. The experimental manipulation of the usage domain consisted of three factors: laboratory set up, task wording and priming task. With regard to the *laboratory set up* for the leisure condition, the lab was set up like a living room, containing a sofa (on which the participant was seated), wooden furniture with travel books, a (switched off) TV set, plants on the window sill, and pictures on the wall. In the work condition, the laboratory contained several desks, a (switched off) computer, a desk lamp, some folders and typical office stationery (stapler, etc.). The *tasks* used in this study were the same in all experimental groups with regard to the interactions they required to be accomplished successfully. Tasks differed however with regard to the framing that was used, with work-specific context presented in the work condition (e.g. plan a meeting in a café with colleagues to discuss a work-related assignment) and leisure-context (e.g. plan a get-together to meet with friends in a café) for the leisure condition. For the *priming task* in the work condition, participants were asked to imagine that they would be working the following two days and to think about what they would have to do during these days. A similar instruction was given in the condition simulating the leisure context.

The experimenter described the purpose of the experiment as testing the usability of a web application for smartphones, giving an overview of the experimental procedure. Participants filled in the PANAS and the questionnaire measuring previous mobile phone experience. The experimenter presented the test device, showed all functions of the web application and explained how to operate it (e.g. choosing categories, home, back, scrolling). Participants completed a practice trial to become familiar with the web application. They were given the opportunity to ask questions, and then instructions about the usage context were provided and participants were asked to start the introspection phase. After one minute of introspection for putting oneself into a specific work or leisure situation, the experimenter informed the participants about the first task. They had five minutes for each task, but were not informed about this time constraint. If the task was not completed after five minutes, the experimenter thanked them and presented them the next task. After the last task, the participants completed the PANAS a second time, then the NASA-TLX, the subjective usability questionnaires (one-item scale and PSSUQ), and finally the manipulation check. The participants were debriefed and could leave their e-mail address to take part in the draw. The duration of a testing session was about 45 min.

3.6 Manipulation check

The manipulation check consisted of a visual analogue scale (0-100; ranging from ‘rather leisure-oriented’ to ‘rather work-oriented’), on which participants judged the situation in which they completed the tasks. Results confirmed a significant impact of the context manipulation, as participants indicated to have experienced the situation signif-

icantly more work-related in the work context condition (M = 56.4; SD = 2.2), compared to the leisure context condition (M = 27.7; SD = 2.4; $t(58)$ = 4.82; $p < 0.001$, Cohen's d = 1.24).

4 Results

Self-reported mobile phone experience, daily mobile usage and gender were entered as covariates into the analysis in order to control for their influence. However, the analysis showed that none of the covariates had a significant influence on the reported findings. Therefore, results of the data analysis without covariates are reported.

4.1 User performance

Task completion rate
Data analysis showed significant differences in the number of completed tasks as a function of the usability manipulation (see Table 1). When operating the system with a delayed response, participants solved significantly fewer tasks (M = 83.3%) than when response time was not delayed (M = 93.3%; F = 5.28; df = 1, 56; $p < 0.05$; η^2_{partial} = .086). Testing context (work vs. leisure) had no effect and there was no interaction between usability and testing context (both $F < 1$).

Task completion time
Overall task completion time differed significantly with regard to the usability manipulation (see Table 1). When operating the system with a delayed response, participants took longer to complete the tasks (corrected by the delay time) than when response time was not delayed ($M_{delayed}$ = 526.7, SD = 145.4; $M_{undelayed}$ = 397, SD = 126.5; F = 13; df = 1, 56; $p < 0.05$; η^2_{partial} = .19). Testing context (work vs. leisure) had no effect and there was no interaction between usability and testing context (both $F < 1$).

Page inspection time
As the data in Table 1 indicates, participants in the delay condition stayed significantly longer on a page (M = 5.84s) than those working with a non-delayed system (M = 5.33s). This difference was statistically significant (F = 4.46; df = 1, 56; $p < 0.05$; η^2_{partial} = .188). With regard to the other independent factors, there was neither a significant effect of testing context ($F < 1$) nor an interaction ($F < 1$).

Efficiency of task completion
An important indicator of user efficiency is determined by the calculation of the ratio of the actual number of user inputs and the optimal number of user inputs. The data in Table 1 indicate overall a medium level of efficiency of about M = 0.4. This efficiency index shows that 40% of the user inputs contributed towards task completion, whereas the remaining inputs did not directly lead to the task goal or were part of a less direct path towards task completion. This indicates that the tasks were reasonably difficult to

solve. As the data in Table 1 suggests, there was little difference between conditions, which was confirmed by analysis of variance (all $F < 1$).

4.2 Subjective ratings

Affective state

For the analysis of the emotional state of the user as a consequence of using the product, a comparison was made between the baseline measurement (i.e. prior to task completion) and a second measurement taken after task completion. This analysis revealed a change in positive affect as a function of SRT. While participants reported an increase in positive affect after task completion when working with a non-delayed system (M = 0.12), lower positive affect was reported when working with a delayed system (M = -0.17; F = 4.67; df = 1, 56; $p < 0.05$; $\eta^2_{\text{partial}} = .077$). Regarding the changes in negative affect, no significant difference was found ($F < 1$). As the data in Table 1 shows, testing context had no effect on the change of positive affect levels and there was no interaction either (both $F < 1$). Equally, there was no effect on the change of negative affect (F = 2.98; df = 1, 56; $p > 0.05$; $\eta^2_{\text{partial}} = .051$), nor was there an interaction ($F < 1$).

Task load

The data for the overall NASA-TLX score are presented in Table 1. While this indicates overall a low task load score, there was generally very little difference between experimental conditions. This was confirmed by analysis of variance, which revealed neither a main effect for the two independent factors nor an interaction between them (all $F < 1$). To evaluate whether any differences could be found at the single item level, a separate analysis of the NASA-TLX items was carried out. Also, this analysis did not show any significant effect.

Perceived usability

The data for perceived usability, as measured by the PSSUQ, are presented in Table 1. Interestingly, the expected effect of SRT did not affect subjective usability evaluations, with ratings being nearly identical in both conditions ($F < 1$). Usability ratings appeared to be higher in the work domain than in the leisure domain but this difference failed to reach significance (F = 2.01; df = 1, 52; $p > 0.05$). No interaction between the two factors was found ($F < 1$). An additional analysis examined the PSSUQ subscales separately but revealed the same pattern of results as for the overall scale. Finally, results for the one-item usability scale indicated an overall rating of M = 52.2 with little differences between the four experimental conditions (all $F < 1$), herewith confirming the results pattern found for the PSSUQ.

Table 1. Measures of user performance, affective state, task load, and perceived usability as a function of usage domain and usability.

	Leisure context		Work context		
	No delay	Delayed	No delay	Delayed	Overall
	Mean (SD)	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*
Task completion rate (%)	96.4 (9.1)	82.8 (21.8)	90.6 (15.5)	83.9 (18.6)	88.3 (17.5)
Task completion time (s)	395.9 (96.5)	534 (168.7)	397.9 (151.2)	518.3 (119.2)	461.8 (150.1)
Page inspection time (s)	5.34 (1.04)	5.93 (1.19)	5.32 (0.75)	5.73 (0.54)	5.58 (0.94)
Efficiency of task completion	0.41 (0.08)	0.39 (0.15)	0.39 (0.13)	0.38 (0.12)	0.39 (0.12)
Affective state (1-5)					
positive affect (Δ: pre - post)	0.09 (0.42)	-0.15 (0.48)	0.14 (0.56)	-0.20 (0.59)	-0.03 (0.53)
negative affect (Δ: pre - post)	-0.29 (2.23)	-1.31 (2.50)	0.44 (3.05)	0.21 (2.08)	-0.25 (2.55)
Task load (1-20)	8.3 (1.5)	7.4 (2.7)	8.0 (3.0)	7.8 (3.0)	7.8 (2.6)
Perceived usability (1-7)	4.7 (0.88)	4.5 (1.33)	4.9 (1.02)	5.0 (0.86)	4.8 (1.03)

Δ: all values represent changes from baseline (pre) to task completion phase (post); a positive value denotes an increase.

4.3 Correlational analysis of data

In addition to the comparisons between experimental conditions using analyses of variance, the size of correlations between variables may provide further insights for a better understanding of the UX construct and the interplay of its different components. This point has notably been addressed by Hornbæk and Law [49] arguing that studies in this domain should report such correlations in order to facilitate interpretation and comparison of outcomes.

Interestingly, correlation coefficients (see table 2) indicate in general rather low correlations between the different UX measures. Significant relationships were found for the different measures of objective performance, whereas performance indicators showed only small correlations with evaluation of perceived usability and task load. In contrast, subjective evaluation of task load was negatively correlated with perceived usability. Furthermore, negative affect showed significant links with perceived task load and task completion rate (c.f. table 2).

Table 2. Correlations between different UX measures (N = 60).

	Task completion time (s)	Efficiency of task completion	Task load	Perceived usability	Positive Affect	Negative affect
Task completion rate (%)	**-.38****	.12	-.23	.16	.14	**-.29***
Task completion time (s)		**-.72****	.19	-.08	-.17	.18
Efficiency of task completion			-.09	.13	.24	-.11
Task load				**-.26***	-.05	**.36****
Perceived usability					.24	-.25
Positive Affect						.03

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

5 Discussion

The aim of the study was to investigate the influence of usage domain on the outcomes of user experience evaluations and whether any such influence would be mediated by poor system usability in the form of SRT delays. The findings showed that, contrary to expectations, usage domain did not have the expected impact, with none of the measures showing differences between domains. In contrast, system response time showed the expected effects on performance and on user emotion whereas, surprisingly, no influence on perceived usability was observed.

Given that context of use has been considered an important determinant of usability [4] and that the two domains of work and leisure have been associated with different perceptions and behaviour [26], we expected that testing a product in one domain would produce differences in usability test results compared to the other domain. The manipulation check showed very clearly that participants perceived the leisure domain differently from the work domain. Despite this successful manipulation of context (involving different usability lab set-ups, domain-specific task instructions, and a priming task), there were no differences in usability test results, neither for performance nor for subjective measures. Although it is important to interpret non-significant results with caution [50], the publication and discussion of such findings is still very important [51–53].

A possible interpretation of this nil-result might be that there is no need for practitioners to test dual-domain products in both usage domains. The domains of work and leisure may not require specific consideration in test set-ups, as long as the relevant use cases are covered in the test. The absence of an interaction between usage domain and

system usability strengthens this argument, suggesting that even under conditions of impaired system usability the work domain provides test results that are no different from the leisure domain. One previous study comparing work and leisure domains also found little difference between these two application domains [13]. However, in that study the usability of the technical device was not manipulated. Taken together, this study and the present work provide first evidence that across a range of conditions (i.e. different levels of product aesthetics and of product usability) the influence of usage domains appears to be of smaller magnitude than expected. The results support Lindroth and Nilsson's [54] claim that environmental aspects of usage context are generally not an important issue in usability testing as long as stationary technology usage is concerned (which was the case in the present study as the smartphone was operated like a desktop device). While Lindroth and Nilsson did not empirically test their proposition, the present work provides first empirical evidence to support it. Additional research (i.e. replication studies) corroborating these results are required however, in order to be able to interpret such nil-results as arguments for practitioners to refrain from considering usage context in UX evaluation.

While these nil-results do not allow us (yet) to make such a decisive statement with regard to the context-dependency of UX evaluation outcomes, the findings have some implications for researchers and practitioners interested in the domain dependence of UX evaluation. The results of this study indicate that the manipulation of the usage context (as suggested in this piece of research) did not show the expected effect. Although the successful manipulation check indicated that a distinction was made by participants in this study, this manipulation showed no influence on UX measures. This may raise some concerns with regard to the extent to which motivational processes associated with the usage context could be appropriately reproduced in the lab. However, it has to be noted that this problem would affect all lab-based UX assessment, independently of the domain. In addition, previous research has shown that lab-based testing often provides similar results compared to conducting tests in the field [55]. In this context, the sample recruited for this study needs to be considered as a limitation since the work context of students may not be fully transferable to salary workers. Therefore it might be worthwhile to address this research question with an additional user sample. Nonetheless do these results provide a strong argument for the need for additional field-based research addressing the influence of usage-context in UX assessment, preferably by making a direct comparison with lab-based data.

While usage domain had little impact on the results of usability testing, a number of effects of poor system usability were found, confirming several of our research hypotheses. First, it emerged that poor system usability had the hypothesised negative effect on task performance. When SRT was delayed, task completion rate was lower and participants spent more time on a page, compared to participants working with a system without delays. These findings are consistent with an extensive body of research showing a negative impact of delayed system response on performance [e.g. 31, 37]. One explanation for this effect is that users adapt their speed of task completion to SRT and work faster when the system responds more promptly [56]. An alternative explanation for longer page inspection times under delayed SRT could be that participants adjusted their strategy, moving from a trial and error approach to a more reflective one, thus

reducing the number of delayed system responses. Previous work has shown that even short SRT delays made participants consider their actions more carefully [38, 57].

Second, poor usability had a negative effect on participants' affective state, consistent with our hypothesis. When working with a delayed system, participants showed a stronger reduction in positive affect than when working with a non-delayed system. This finding is consistent with an extensive body of research, showing negative effects of delayed SRT on various aspects of affective states, such as frustration, anxiety, stress and impatience [e.g. 38–40]. The present study adds to these findings by showing that such effects on emotion may occur, even if such SRT delays are short.

Third, although poor system usability had a negative effect on performance and affected participants' emotional state, no such effects were found for perceived usability. This observation is of particular interest since other work found a substantial positive association between performance and preference [58]. While users generally provide a more positive evaluation when systems are more usable, Nielsen and Levy [58] also cited some cases in their meta-analysis, in which users preferred systems, with which they performed worse. These systems, however, had rather short SRT delays and performance impairments did not reach critical levels. The magnitude of the delay in our study might have been below that critical level and therefore did not have an effect on perceived usability. An alternative explanation for the observed finding is that participants did perceive such delays but the SRT delay was not associated with the application but with the server from which the pages were downloaded. Similar observations were made in other work where users of internet-based software attributed the cause of delayed response to internet connection rather than the software itself [59]. Overall, although we employed rather short SRT delays, most of our hypotheses were confirmed, which highlights the importance of paying attention to even short delays during system design as it may affect performance and user emotion.

The correlational analysis of the different measures of the UX construct revealed in general rather low correlations. Objective performance measures, subjective evaluations of usability and affective states as important aspects describing an experience episode seem to be rather independent dimensions and hence should be assessed and reported separately when UX is the scope of measurement. Hence, a combination into a single UX-score seems not to be useful. These findings are in line with results of previous studies indicating that measures of user performance often have only a weak link to subjective measures of usability [49, 57, see also 60]. Radar charts might represent a useful and easy to understand way to display results of a holistic UX evaluation consisting of different dimensions or facets. Results of this study could represented as suggested in Figure 2 with regard to the evaluation of the two different versions of the prototype.

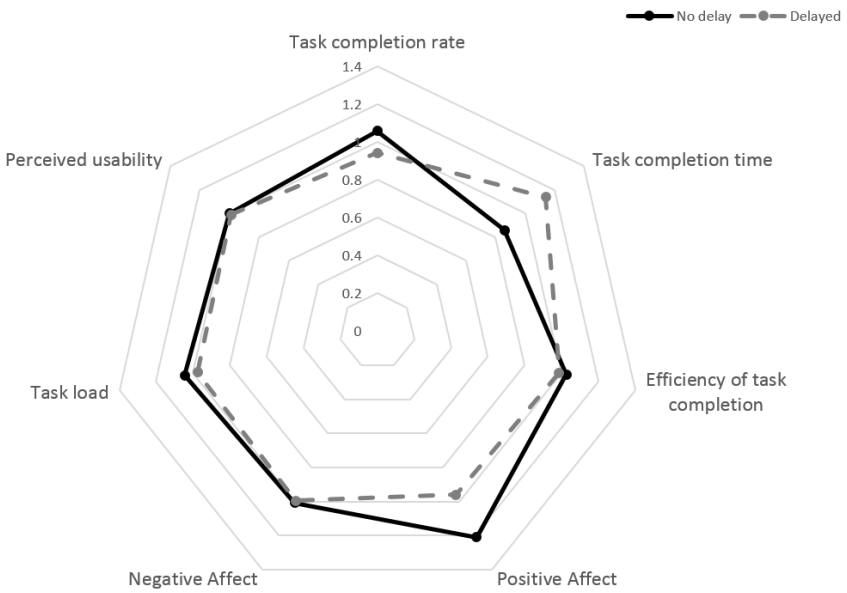

Fig. 2. Presentation of the results of UX evaluations of the two versions of the mobile app as radar diagram

The findings presented have several implications for research and practice. First, the findings provide first evidence that results of a usability test can be transferred between the work and the leisure domain. This would facilitate usability testing of dual-domain products for practitioners since several testing contexts would not have to be covered so that they only need to ensure that the relevant tasks are included in the test set-up. Second, practitioners and researchers interested in context-dependent UX evaluation should address this issue in field-based research. Third, even rather short delays in SRT can have an effect on performance as well as on user emotion, suggesting that careful consideration should be given to SRT in product design and evaluation.

References

1. Law, E., Roto, V., Vermeeren, A.P.O.S., Kort, J., Hassenzahl, M.: Towards a Shared Definition of User Experience. In: CHI ’08 Extended Abstracts on Human Factors in Computing Systems. pp. 2395–2398. ACM, New York, NY, USA (2008). https://doi.org/10.1145/1358628.1358693.
2. Bevan, N., Macleod, M.: Usability measurement in context. Behaviour & Information Technology. 13, 132–145 (1994). https://doi.org/10.1080/01449299408914592.

3. ISO 9241-210:2010: Ergonomics of human-system interaction -- Part 210: Human-centred design for interactive systems, http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/20/52075.html, (2010).
4. Maguire, M.: Context of Use within usability activities. International Journal of Human-Computer Studies. 55, 453–483 (2001). https://doi.org/10.1006/ijhc.2001.0486.
5. Alonso-Ríos, D., Vázquez-García, A., Mosqueira-Rey, E., Moret-Bonillo, V.: A Context-of-Use Taxonomy for Usability Studies. International Journal of Human–Computer Interaction. 26, 941–970 (2010). https://doi.org/10.1080/10447318.2010.502099.
6. Bargas-Avila, J.A., Hornbaek, K.: Old Wine in New Bottles or Novel Challenges: A Critical Analysis of Empirical Studies of User Experience. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 2689–2698. ACM, New York, NY, USA (2011). https://doi.org/10.1145/1978942.1979336.
7. Nielsen, C.M., Overgaard, M., Pedersen, M.B., Stage, J., Stenild, S.: It's Worth the Hassle!: The Added Value of Evaluating the Usability of Mobile Systems in the Field. In: Proceedings of the 4th Nordic Conference on Human-computer Interaction: Changing Roles. pp. 272–280. ACM, New York, NY, USA (2006). https://doi.org/10.1145/1182475.1182504.
8. Sonderegger, A., Sauer, J.: The influence of laboratory set-up in usability tests: effects on user performance, subjective ratings and physiological measures. Ergonomics. 52, 1350–1361 (2009). https://doi.org/10.1080/00140130903067797.
9. Uebelbacher, A., Sonderegger, A., Sauer, J.: Effects of Perceived Prototype Fidelity in Usability Testing under Different Conditions of Observer Presence. Interact Comput. 25, 91–101 (2013). https://doi.org/10.1093/iwc/iws002.
10. Harris, E., Weinberg, J., Thomas, S., Gaeslin, D.: Effects of social facilitation and electronic monitoring on usability testing. In: Proc. Usability Prof. Assoc. Conf. pp. 1–8 (2005).
11. Grudin, J.: Three faces of human-computer interaction. IEEE Annals of the History of Computing. 27, 46–62 (2005). https://doi.org/10.1109/MAHC.2005.67.
12. den Buurman, R.: User-centred design of smart products. Ergonomics. 40, 1159–1169 (1997). https://doi.org/10.1080/001401397187676.
13. Sonderegger, A., Uebelbacher, A., Pugliese, M., Sauer, J.: The Influence of Aesthetics in Usability Testing: The Case of Dual-domain Products. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 21–30. ACM, New York, NY, USA (2014). https://doi.org/10.1145/2556288.2557419.
14. Carroll, J.M.: Community computing as human-computer interaction. Behaviour & Information Technology. 20, 307–314 (2001). https://doi.org/10.1080/01449290110078941.
15. Tractinsky, N.: The Usability Construct: A Dead End? Human–Computer Interaction. 33, 131–177 (2018). https://doi.org/10.1080/07370024.2017.1298038.
16. Forlizzi, J., Battarbee, K.: Understanding Experience in Interactive Systems. In: Proceedings of the 5th Conference on Designing Interactive Systems: Processes,

Practices, Methods, and Techniques. pp. 261–268. ACM, New York, NY, USA (2004). https://doi.org/10.1145/1013115.1013152.
17. Wright, P., McCarthy, J., Meekison, L.: Making Sense of Experience. In: Blythe, M.A., Overbeeke, K., Monk, A.F., and Wright, P.C. (eds.) Funology. pp. 43–53. Springer Netherlands (2003). https://doi.org/10.1007/1-4020-2967-5_5.
18. Desmet, P., Hekkert, P.: Framework of product experience. International journal of design, 1(1)2007. (2007).
19. Hassenzahl, M.: User Experience (UX): Towards an Experiential Perspective on Product Quality. In: Proceedings of the 20th Conference on L'Interaction Homme-Machine. pp. 11–15. ACM, New York, NY, USA (2008). https://doi.org/10.1145/1512714.1512717.
20. Law, E.L.-C., van Schaik, P., Roto, V.: Attitudes towards user experience (UX) measurement. International Journal of Human-Computer Studies. 72, 526–541 (2014).
21. Minge, M., Thüring, M., Wagner, I., Kuhr, C.V.: The meCUE Questionnaire: A Modular Tool for Measuring User Experience. In: Soares, M., Falcão, C., and Ahram, T.Z. (eds.) Advances in Ergonomics Modeling, Usability & Special Populations. pp. 115–128. Springer International Publishing (2017). https://doi.org/10.1007/978-3-319-41685-4_11.
22. Bevan, N.: Classifying and selecting UX and usability measures. In: International Workshop on Meaningful Measures: Valid Useful User Experience Measurement. pp. 13–18. Institute of Research in Informatics of Toulouse (IRIT), Toulouse, France (2008).
23. Thüring, M., Mahlke, S.: Usability, aesthetics and emotions in human–technology interaction. International Journal of Psychology. 42, 253–264 (2007). https://doi.org/10.1080/00207590701396674.
24. Mandryk, R.L., Inkpen, K.M., Calvert, T.W.: Using psychophysiological techniques to measure user experience with entertainment technologies. Behaviour & Information Technology. 25, 141–158 (2006). https://doi.org/10.1080/01449290500331156.
25. Beatty, J.E., Torbert, W.R.: The False Duality of Work and Leisure. Journal of Management Inquiry. 12, 239–252 (2003). https://doi.org/10.1177/1056492603256340.
26. Rheinberg, F., Manig, Y., Kliegl, R., Engeser, S., Vollmeyer, R.: Flow bei der Arbeit, doch Glück in der Freizeit. Zeitschrift für Arbeits- und Organisationspsychologie A&O. 51, 105–115 (2007). https://doi.org/10.1026/0932-4089.51.3.105.
27. Tinsley, H.E., Hinson, J.A., Tinsley, D.J., Holt, M.S.: Attributes of leisure and work experiences. Journal of Counseling Psychology. 40, 447–455 (1993). https://doi.org/10.1037/0022-0167.40.4.447.
28. Shneiderman, B.: Response time and display rate in human performance with computers. ACM Computing Surveys (CSUR). 16, 265–285 (1984).
29. Szameitat, A.J., Rummel, J., Szameitat, D.P., Sterr, A.: Behavioral and emotional consequences of brief delays in human–computer interaction. International Journal

of Human-Computer Studies. 67, 561–570 (2009). https://doi.org/10.1016/j.ijhcs.2009.02.004.
30. Roto, V., Oulasvirta, A.: Need for Non-visual Feedback with Long Response Times in Mobile HCI. In: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web. pp. 775–781. ACM, New York, NY, USA (2005). https://doi.org/10.1145/1062745.1062747.
31. Barber, R.E., Lucas Jr., H.C.: System Response Time, Operator Productivity and Job Satisfaction. Social Science Research Network, Rochester, NY (1983).
32. Dellaert, B.G.C., Kahn, B.E.: How tolerable is delay?: Consumers' evaluations of internet web sites after waiting. Journal of Interactive Marketing. 13, 41–54 (1999). https://doi.org/10.1002/(SICI)1520-6653(199924)13:1<41::AID-DIR4>3.0.CO;2-S.
33. Hoxmeier, J.A., DiCesare, C.: System response time and user satisfaction: An experimental study of browser-based applications. AMCIS 2000 Proceedings. 140–145 (2000).
34. Rushinek, A., Rushinek, S.F.: What Makes Users Happy? Commun. ACM. 29, 594–598 (1986). https://doi.org/10.1145/6138.6140.
35. Ramsay, J., Barbesi, A., Preece, J.: A psychological investigation of long retrieval times on the World Wide Web. Interacting with Computers. 10, 77–86 (1998). https://doi.org/10.1016/S0953-5438(97)00019-2.
36. Butler, T.W.: Computer Response Time and User Performance. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 58–62. ACM, New York, NY, USA (1983). https://doi.org/10.1145/800045.801581.
37. Galletta, D.F., Henry, R., McCoy, S., Polak, P.: Web Site Delays: How Tolerant are Users? Journal of the Association for Information Systems. 5, (2004).
38. Guynes, J.L.: Impact of System Response Time on State Anxiety. Commun. ACM. 31, 342–347 (1988). https://doi.org/10.1145/42392.42402.
39. Polkosky, M.D., Lewis, J.R.: Effect of Auditory Waiting Cues on Time Estimation in Speech Recognition Telephony Applications. International Journal of Human–Computer Interaction. 14, 423–446 (2002). https://doi.org/10.1080/10447318.2002.9669128.
40. Selvidge, P.R., Chaparro, B.S., Bender, G.T.: The world wide wait: effects of delays on user performance. International Journal of Industrial Ergonomics. 29, 15–20 (2002). https://doi.org/10.1016/S0169-8141(01)00045-2.
41. Trimmel, M., Meixner-Pendleton, M., Haring, S.: Stress Response Caused by System Response Time when Searching for Information on the Internet. Hum Factors. 45, 615–622 (2003). https://doi.org/10.1518/hfes.45.4.615.27084.
42. Jacko, J.A., Sears, A., Borella, M.S.: The effect of network delay and media on user perceptions of web resources. Behaviour & Information Technology. 19, 427–439 (2000). https://doi.org/10.1080/014492900750052688.
43. Hui, M.K., Tse, D.K.: What to Tell Consumers in Waits of Different Lengths: An Integrative Model of Service Evaluation. Journal of Marketing. 60, 81–90 (1996). https://doi.org/10.1177/002224299606000206.

44. Branaghan, R.J., Sanchez, C.A.: Feedback Preferences and Impressions of Waiting. Hum Factors. 51, 528–538 (2009). https://doi.org/10.1177/0018720809345684.
45. Watson, D., Clark, L.A., Tellegen, A.: Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology. 54, 1063–1070 (1988). https://doi.org/10.1037/0022-3514.54.6.1063.
46. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Meshkati, P.A.H. and N. (ed.) Advances in Psychology. pp. 139–183. North-Holland (1988).
47. Christophersen, T., Konradt, U.: Reliability, validity, and sensitivity of a single-item measure of online store usability. International Journal of Human-Computer Studies. 69, 269–280 (2011). https://doi.org/10.1016/j.ijhcs.2010.10.005.
48. Lewis, J.R.: IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. International Journal of Human–Computer Interaction. 7, 57–78 (1995). https://doi.org/10.1080/10447319509526110.
49. Hornbæk, K., Law, E.L.-C.: Meta-analysis of Correlations Among Usability Measures. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 617–626. ACM, New York, NY, USA (2007). https://doi.org/10.1145/1240624.1240722.
50. Altman, D.G., Bland, J.M.: Statistics notes: Absence of evidence is not evidence of absence. BMJ. 311, 485 (1995). https://doi.org/10.1136/bmj.311.7003.485.
51. Chalmers, L.: Underreporting Research Is Scientific Misconduct. JAMA. 263, 1405–1408 (1990). https://doi.org/10.1001/jama.1990.03440100121018.
52. Fanelli, D.: Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data. PLOS ONE. 5, e10271 (2010). https://doi.org/10.1371/journal.pone.0010271.
53. Stern, J.M., Simes, R.J.: Publication bias: evidence of delayed publication in a cohort study of clinical research projects. BMJ. 315, 640–645 (1997). https://doi.org/10.1136/bmj.315.7109.640.
54. Lindroth, T., Nilsson, S., Rasmussen, P.: Mobile usability—rigour meets relevance when usability goes mobile. In: Bjørnestad, S., Moe, R.E., Mørch, A.I., and Opdahl, A.L. (eds.) Proceedings of the 24th information systems research seminar in Scandinavia (IRIS'24). pp. 641–654. IRIS Association, Ulvik (2001).
55. Kjeldskov, J., Stage, J.: New techniques for usability evaluation of mobile systems. International Journal of Human-Computer Studies. 60, 599–620 (2004). https://doi.org/10.1016/j.ijhcs.2003.11.001.
56. Boucsein, W.: Forty Years of Research on System Response Times – What Did We Learn from It? In: Schlick, C.M. (ed.) Industrial Engineering and Ergonomics. pp. 575–593. Springer Berlin Heidelberg (2009).
57. Teal, S.L., Rudnicky, A.I.: A Performance Model of System Delay and User Strategy Selection. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 295–305. ACM, New York, NY, USA (1992). https://doi.org/10.1145/142750.142818.

58. Nielsen, J., Levy, J.: Measuring Usability: Preference vs. Performance. Commun. ACM. 37, 66–75 (1994). https://doi.org/10.1145/175276.175282.
59. Rose, G.M., Meuter, M.L., Curran, J.M.: On-line waiting: The role of download time and other important predictors on attitude toward e-retailers. Psychology & Marketing. 22, 127–151 (2005). https://doi.org/10.1002/mar.20051.
60. Sonderegger, A., Sauer, J.: The role of non-visual aesthetics in consumer product evaluation. International Journal of Human-Computer Studies. 84, 19–32 (2015). https://doi.org/10.1016/j.ijhcs.2015.05.011.